OPEN LETTER

Genomic variant sharing: a position statement [version 1; peer review: 1 approved, 1 approved with reservations]

Caroline F. Wright¹, James S. Ware², Anneke M. Lucassen³, Alison Hall⁴, Anna Middleton⁵,⁶, Nazneen Rahman⁷, Sian Ellard¹, Helen V. Firth⁸,⁹

¹Institute of Biomedical and Clinical Science, University of Exeter, Exeter, UK
²National Heart and Lung Institute, Imperial Centre for Translational and Experimental Medicine, London, UK
³Department of Clinical Ethics and Law, Faculty of Medicine, University of Southampton, Southampton, UK
⁴PHG Foundation, Cambridge, UK
⁵Faculty of Education, University of Cambridge, Cambridge, UK
⁶Connecting Science, Wellcome Genome Campus, Cambridge, UK
⁷Division of Genetics and Epidemiology, Institute of Cancer Research, London, UK
⁸Department of Clinical Genetics, University of Cambridge Addenbrooke's Hospital Cambridge, Cambridge, UK
⁹Wellcome Trust Sanger Institute, Cambridge, UK

Abstract
Sharing de-identified genetic variant data is essential for the practice of genomic medicine and is demonstrably beneficial to patients. Robust genetic diagnoses that inform medical management cannot be made accurately without reference to genetic test results from other patients, as well as population controls. Errors in this process can result in delayed, missed or erroneous diagnoses, leading to inappropriate or missed medical interventions for the patient and their family. The benefits of sharing individual genetic variants, and the harms of not sharing them, are numerous and well-established. Databases and mechanisms already exist to facilitate deposition and sharing of pseudonomised genetic variants, but clarity and transparency around best practice is needed to encourage widespread use, prevent inconsistencies between different communities, maximise individual privacy and ensure public trust. We therefore recommend that widespread sharing of a small number of individual genetic variants associated with limited clinical information should become standard practice in genomic medicine. Information robustly linking genetic variants with specific conditions is fundamental biological knowledge, not personal information, and therefore should not require consent to share. For additional case-level detail about individual patients or more extensive genomic information, which is often essential for clinical interpretation, it may be more appropriate to use a controlled-access model for data sharing, with the ultimate aim of making as much information as open and de-identified as possible with appropriate consent.

Keywords
medical genomics, variant, data sharing, data ethics
Recommendations

1. Open and widespread sharing of plausibly causal genetic variants with high-level phenotypes should be routine clinical practice and should not be dependent upon consent from individual patients.

2. A single genetic variant is not personally identifiable information; however, it is good practice to maintain a cryptic link to the laboratory or clinical service that shared the genetic data so that clinical follow-up remains possible should knowledge of the implications of a variant change.

3. Disclosing detailed case-level clinical detail, larger variant sets or genome-wide data outside of the patient’s clinical team may be crucial for variant interpretation or clinical management, but requires explicit consent to share openly.

Introduction

Making an accurate diagnosis is the cornerstone of good medical practice, essential for determining prognosis, guiding treatment and informing patient management. Across all medical specialties, the interpretation of diagnostic test results relies upon knowledge of what is ‘normal’ in the population versus what ‘disease’ looks like. This knowledge relies upon sharing test results from previous patients and population controls. Without such data, the sensitivity and specificity of the test is unknown, its clinical utility is questionable, and its continued use may be harmful.

Genomic medicine is no exception to this rule, but determining what constitutes ‘normal’ and ‘disease’ can be extremely complicated and arguably the need for ongoing data sharing is even greater than in other branches of medicine. Increasingly, clinical testing will rely on genome-wide sequencing, rather than targeted single-gene testing, and the enormous amount of normal variation in every genome means that interpreting the results from one person’s genome requires knowledge of thousands of other genomes across different populations. Despite ongoing efforts to sequence large cohorts, every genome examined contains novel changes not previously seen. For diseases with a substantial genetic component, caused by a specific rare variant in an individual’s genome, determining which variants are responsible for disease—and which are simply incidental—is an enormous challenge. The only way to meet that challenge is by sharing data on individual variants with associated high-level phenotype terms.

Advantages of sharing genetic variant data

The main purpose of sharing individual genetic variants is to improve the diagnostic accuracy of genetic testing; the main data processors are clinicians and clinical scientists, and the main beneficiaries are patients. Within this context, there are many benefits of sharing individual genetic variants associated with specific conditions:

1. **Making accurate and safe diagnoses.** Genetic testing often benefits the individual patient undergoing testing, whose diagnosis can be accurately determined and prognosis further refined. All such genetic testing is dependent on being able to compare the variant of interest to variants from thousands of other people (via a database that is accessed by the scientist or clinician doing the analysis); at a minimum, this variant comparison is necessary to characterise and usually exclude variants that are relatively common in the general population. Numerous examples exist where making a successful genetic diagnosis has only been possible as a result of being able to access variant and phenotype data from other individuals undergoing testing, and many new genetic causes of disease have been uncovered this way. While most of the published cases are clinician-led, there are an increasing number of patient-led examples of variant sharing that have also catalysed the formation of disease-specific patient support groups and created new avenues of research.

2. **More effective disease management and precision medicine.** In some cases, an accurate genetic diagnosis leads to specific targeted therapies that can more effectively treat disease, or, in rare cases, may even reverse or prevent disease. As a result of variant sharing, individuals may also be recruited to clinical trials that are tailored to their specific genotype, offering the potential for therapy where none currently exists.

In addition, new fundamental biological insights from genetic studies may identify novel targets for future therapies.

3. **Accurate advice for family members.** Due to the shared familial nature of most genetic changes, the benefits of making a robust genetic diagnosis may be cascaded out to biological relatives and have a profound impact on both existing and future generations. Given the value of making a genetic diagnosis across a plethora of different (mostly individually rare) diseases, a strong argument can be made that individuals have a moral duty to help family members and other similar patients by allowing information derived from them to be shared if it is easy to do so.

4. **Improved understanding of genetic disease.** There are also wider benefits to the community, including patients, clinicians and researchers across the globe, who are trying to understand and treat the causes of disease. Reports of new gene-disease associations, and sharing of variant-level information to discern which specific variants within each gene are pathogenic or benign, or carry some degree of risk, are critical to advance our understanding of genetic disease. Moreover, sharing variants together with phenotype and age will allow an evolving understanding of penetrance, improving interpretation of both diagnostic and predictive testing.

Disadvantages of not sharing genetic variant data

There is a substantial opportunity cost to not sharing clinically-oriented data that could otherwise be used to accelerate medical
progress. The harms of not sharing individual genetic variants are well established and include delayed, missed and erroneous diagnoses, leading to inappropriate care and sometimes litigation. (See Box 1 and Box 2 for examples where variant sharing had a direct impact on clinical care.) Due to the familial nature of genetics, any diagnostic mistakes can easily be compounded by cascading erroneous information out to family members, thus multiplying the harms. Furthermore, without data sharing, research progress would be impeded, and the growing genomics knowledgebase—upon which the promise of personalised medicine is based—will stagnate. Historical mistakes that exist in public variant databases can never be fixed without an influx of new data to allow reclassification of variants, without which misdiagnoses and errors in predictive algorithms will continue. Individual organisations that actively maintain private genetic variant databases, such as commercial companies that do not share variant information for proprietary reasons, are thus inhibiting diagnoses for other patients and undermining public health efforts in this area.

Perceived harms of sharing genetic variant data
We have not been able to find any evidence that sharing data relating to individual genetic variants in the context of clinical applications causes harm. Nonetheless, perceived harms include re-identification of individuals across different datasets, loss of security of associated medical information (about the individual or their relatives), and the maleficent misuse of data. Early fears relating to genetic discrimination and the impact of genetic data on insurance premiums have proven to be largely unfounded in the UK and many other countries, thanks in part to genetic non-discrimination legislation and the ongoing Concordat on Genetics and Insurance. Identification of an individual through knowledge of their genetic variant(s) is now perhaps the main concern. Although it is never possible to guarantee anonymity, individual genetic variants—even very rare ones—are not uniquely identifying, and re-identification would require an intimate knowledge of the individual’s genotype or phenotype together with some information to trace that genotype/phenotype to a specific person. In practice, only an individual patient or their clinician would easily be able to re-identify themselves from a specific variant, neither of which would constitute a breach of confidentiality. A related concern is the perception that all genetic data are personal and therefore inherently sensitive, which stems from conflating genome-wide data with individual genetic variants. All possible genetic variants (of which there are many billions) can easily be computed and their predicted consequences stored; if one particular variant is then found in a patient, it makes no sense for any of the existing information about that variant to then become personal or unique. The only new information is simply that the variant has now been seen in a particular disease case and not in controls.

Finding a balance
In our view, the definite and provable harms of not sharing genetic data outweigh the potential and largely hypothetical harms of sharing, a view that is corroborated by several recent litigation cases and supported by several large opinion surveys. Some empirical research has shown that patients and research participants support widespread data sharing and believe that the positive consequences outweigh the potential negatives. Recognising these benefits, 13 European countries have recently signed a declaration for delivering cross-border access to their genomic information. Nonetheless, in our

---

**Box 1. Example 1: The hazard of variant over-interpretation.**

In the early 2000’s, a routine scan from a woman in her second trimester of pregnancy showed increased signal in the fetal bowel. This can be a sign of a chromosomal anomaly, viral infection or cystic fibrosis (CF) so an amniocentesis was offered. DNA analysis showed the fetus carried two CFTR variants that were said to be pathogenic. The parents were counselled that their baby would be affected by CF. They elected to continue the pregnancy.

After birth, the child was started on prophylactic antibiotics, twice daily physiotherapy, regular nebulisers and pancreatic supplements. Years later, the child was referred to the genetics clinic for review because the disease seemed unusually mild.

**The clinical geneticist told the family that the status of one mutation had changed in the CFTR2 database and this combination was no longer thought to cause cystic fibrosis.**

As a direct consequence of this change in variant interpretation, the child’s prognosis changed from a life-limiting disorder to one of near-normal life expectancy and the day-to-day life of the child was transformed. The intensive regime of care was substantially reduced.

**Box 2. Example 2: The need for population-specific variation data.**

A middle-aged Turkish man was referred to clinical genetics because he had colorectal cancer and numerous polyps were discovered at surgery. A homoygous variant in MUTYH was identified and reported to be of “unknown significance” in the diagnostic laboratory report. Biallelic MUTYH mutations cause MUTYH-associated polyposis (MAP), a recessive syndrome consistent with the diagnosis. Specific mutations are found at different frequencies in different populations.

Evaluation of available databases revealed that the variant had been identified once before in a patient with colon cancer and polyposis. Notably this second patient was also Turkish. No functional data were available and in silico analyses were inconclusive. The variant is extremely rare; present in only 7 individuals, all of South or East Asian origin, in the Exome Aggregation Data set of 61,486 individuals. However, no Turkish samples are listed as contributing to any of these datasets and no MUTYH or exome data from the general Turkish population is available.

Thus it is unclear whether this MUTYH variant is a pathogenic Turkish founder mutation or a non-pathogenic variant that is particularly prevalent in the Turkish population, but rare/absent in other populations. This lack of clarity presents significant clinical challenges in managing the patient and his relatives. Sharing data generated in laboratories worldwide and across more ethnic groups would provide information to differentiate between these options and would allow clear classification of this and many other variants and reduce the potential for health disparities.
increasingly data-aware society, there is a perception that data sharing is inherently risk\textsuperscript{2}. A balance must therefore be struck between sharing sufficient data to reap the benefits, but only as much data as is needed to avoid the potential (perceived and actual) harms.

We have previously proposed a principle of proportionality in genetic data sharing, that balances the depth of data shared with the breadth of sharing\textsuperscript{3,4}. With any dataset, decisions must be made about what specifically to share and how widely to share it. Many of the clinical benefits of data sharing in genetics can be realised by sharing a tiny subset of de-identified genetic variants, together with limited medical data, rather than necessarily whole genomes. This principle is in accordance with data privacy laws such as the new European General Data Protection Regulation (GDPR), which mandates that stored data are “adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed”\textsuperscript{4}.

The specifics of implementation are critical, and agreeing standards for sharing variants and associated clinical data is essential. Specific data elements for sharing individual genetic variants have been outlined previously\textsuperscript{5,6} and include (see Table 1):

1. a standardised genetic description of the variant(s), including Human Genome Variation Society (HGVS) nomenclature and genomic coordinates of the variant;
2. a standardised clinical description of the clinical features in the patient using appropriately controlled vocabulary/ontology;
3. the inheritance pattern of the disease (e.g. dominant/recessive);
4. the clinical significance and summary of evidence upon which that assertion was based; and
5. a cryptic link to the laboratory or clinical service that submitted the data (to enable further information to be requested and avoid data duplication).

We recommend that openly sharing genetic variant data at this level should be routine practice. No personal identifiers should be openly shared (e.g. name, hospital IDs, geographical location, etc), and only the minimal genetic and clinical information required (as outlined in the five points above) to assist with interpreting a similar variant should be included. We recommend a cryptic link to the individual case-level data is maintained in a de-identified fashion via the laboratory or clinical service that submitted the data, that may obscure its geographical location by deposition via another platform, to enable clinical follow-up if needed. Linking basic clinical information with information about genetic variation is crucial for supporting variant interpretation and aiding diagnoses. However, as with more extensive genome-wide data, or genomic risk scores, different levels of clinical detail will require different modes of sharing, i.e. open versus controlled access. Additional phenotype detail enables a clinical genomics team to assess the strength of the diagnostic claim and evaluate the evidence for a purported diagnosis. Including this detailed clinical information with a genetic test result avoids potential attrition, where individual clinicians need to go back to the original data generator to obtain sufficient information with which to make a diagnosis in their patient.

A flexible platform with broad international sharing of variant data together with national/local sharing of more granular phenotypic data would enable both needs to be addressed. Numerous

---

**Table 1. Example of genomic variant sharing.**

<table>
<thead>
<tr>
<th>Variant 1</th>
<th>Variant 2</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Variant</strong></td>
<td>Standardised description of variant, including</td>
</tr>
<tr>
<td></td>
<td>genomic coordinates</td>
</tr>
<tr>
<td><strong>Gene</strong></td>
<td>e.g. MYH7</td>
</tr>
<tr>
<td><strong>Genotype</strong></td>
<td>Heterozygous</td>
</tr>
<tr>
<td><strong>Phenotype</strong></td>
<td>Hypertrophic cardiomyopathy</td>
</tr>
<tr>
<td>ACMG/AMP variant-level evidence\textsuperscript{6}</td>
<td>PS1 – a different variant at the same position</td>
</tr>
<tr>
<td></td>
<td>PS1 – a different variant at the same position</td>
</tr>
<tr>
<td></td>
<td>previously been established to be pathogenic</td>
</tr>
<tr>
<td></td>
<td>PM1 – occurs in the head of the protein (a</td>
</tr>
<tr>
<td></td>
<td>functional domain with high probability</td>
</tr>
<tr>
<td>interpretation (based on</td>
<td>pathogenicity)</td>
</tr>
<tr>
<td>public data)</td>
<td>PM2 – absent from the general population</td>
</tr>
<tr>
<td></td>
<td>PP3 – computational evidence suggests</td>
</tr>
<tr>
<td>aggregated case-level</td>
<td>deleterious effect on gene product</td>
</tr>
<tr>
<td>evidence\textsuperscript{8}</td>
<td>Lab A – variant observed in 23,000 patients</td>
</tr>
<tr>
<td></td>
<td>sequenced</td>
</tr>
<tr>
<td></td>
<td>Lab B – 2/4,000</td>
</tr>
<tr>
<td></td>
<td>Lab C – 1/3,000</td>
</tr>
<tr>
<td><strong>Interpretation</strong></td>
<td>Likely pathogenic</td>
</tr>
<tr>
<td><strong>Interpretation</strong></td>
<td>Variant of uncertain significance</td>
</tr>
<tr>
<td><strong>Interpreted case-level</strong></td>
<td>Observed in 1/10,000 individuals referred with</td>
</tr>
<tr>
<td>evidence**</td>
<td>diagnosis of HCM</td>
</tr>
<tr>
<td><strong>Interpreted with</strong></td>
<td>Likely pathogenic</td>
</tr>
<tr>
<td><strong>variant sharing</strong></td>
<td>Likely pathogenic</td>
</tr>
</tbody>
</table>

---

Wellcome Open Research 2019, 4:22 Last updated: 17 JUL 2019
databases already exist for collating and sharing genetic information, which may have differing requirements for data deposition and thus offer different advantages and disadvantages. For example, US-based ClinVar is perhaps the leading genetic variant deposition database, with >600,000 open access variants assayed primarily through clinical genetic testing services, but most variants have only very limited or no clinical information and no supporting evidence associated with them. UK-based DECIPHER contains detailed case-level clinical data associated with >65,000 variants, but uses a tiered access model whereby around half the cases are completely open access and half are only accessible to members of closed groups. DECIPHER and many other databases are part of Matchmaker Exchange, which enables data sharing and case-matching across separate and otherwise potentially siloed genetic datasets.

Establishing good practice
Uncertainty about what are permissible types of genetic variant sharing and when explicit consent is required means that current data sharing practices across regional genetics centres are highly variable. The inclusion of genetic data within Article 9 of the European GDPR, “Processing of special categories of personal data”, has created further confusion about the legality of sharing individual variants. There is therefore a need to establish and agree best practice for data sharing within genomic medicine, to avoid inconsistent practices across different regions, communities and jurisdictions, and ensure transparency and consistency when speaking to patients. Genetic variant data of the sort described above does not meet a recently proposed Data Sharing Privacy Test, as the data is neither inherently sensitive nor uniquely identifying. Within the UK, the National Data Guardian has stated that “the duty to share information can be as important as the duty to protect patient confidentiality”, a principle that applies to all data generated across the National Health Service. The American College of Medical Genetics and Genomics recently published a position statement in 2017 that “laboratory and clinical genomic data sharing is crucial to improving genetic health care”. However, genomic medicine is inherently a global enterprise, so more countries need to follow suit. The approach to data sharing espoused by the Global Alliance for Genomics and Health is rooted in international human rights legislation, focussing on our ‘solidarity rights’ to genomic information and emphasising the social good that can derive from appropriate data sharing. The handful of patients with the same rare diagnosis may be scattered across different countries, and are therefore best served when data are shared as openly and as widely as possible. Patients across the globe currently benefit from shared data and derived knowledge in databases such as ClinVar and DECIPHER, and services that are not currently sharing their clinical data owe a substantial data debt and risk perpetuating current data biases.

Separate consent should not be required for individual variant sharing
Clinical experience suggests that most patients are keen for their variant data and associated phenotypes to be shared when the risks and benefits are explained to them. Indeed, discussion of the need for data sharing should be so integral in the discussion about genomic diagnosis that separate consent should not be required. Variants of uncertain significance are regularly generated from genome-wide testing and can most easily be resolved through being able to access and explore the context in which such variants have been observed elsewhere (which depends on data sharing, see Figure 1). In addition to variants from current and future patients, enormous swaths of legacy data exist from

![Figure 1](https://example.com/figure1.png)

**Figure 1.** Global open variant sharing enables robust diagnoses to be made as quickly as possible; facilitating sharing of detailed case-level information also informs clinical management and aids diagnosis in complex cases.
decades of patients who have undergone genetic testing, some of whom are no longer alive, and most of whom are no longer in touch with their clinicians. Sharing variants from these tests could potentially benefit many thousands of patients, and poses little or no harm to the data subjects.

Rather than considering ownership of data as the route to determine what can be done with it, examining who controls access to the data is perhaps a more plausible solution than entering into ownership debates\(^\text{96}\). Individuals have a right to control access to data relating to them, but we argue that when it also pertains to others—as is the case for genetic variants—individual rights of veto should be limited to the most unusual situations. A link between a particular genetic variant and clinical features of a disease is not personal information any more than the link between high blood cholesterol and heart disease, for example. We therefore propose that patient consent should not be required to share data on individual genetic variants, with minimal clinical information sufficient to provide other clinical professionals with a reasonable overview of the case(s) to provide appropriate medical care\(^\text{111}\). Agreeing this principle of “clinical de-identified variant-level sharing”\(^\text{112}\), would remove the onus from data generators to ensure that they have the appropriate written consents and permissions in place, which can be extremely daunting when data may ultimately end up in multiple different databases, and replace it with an unambiguous policy that is clear and transparent for both data generators and data subjects. In addition, for many rare disorders, we suggest that more detailed case-specific information generated within a particular healthcare system should initially remain within that healthcare system, sensitive to the quirks of each individual regulatory regime, but with the aim of eventual open data sharing following discussion with the patient and subject to their explicit consent.

Conclusions

All genetic interpretation is fundamentally dependent upon data sharing, since it is impossible to demonstrate an association between a particular rare genetic change and a disease with an “N-of-one”. Therefore, sharing genetic variant data—albeit aggregated at some level and de-identified as far as possible—is inseparable from the practice of genomic medicine. Clinicians cannot treat patients appropriately if they cannot compare their patient’s data with data from healthy populations and other patients and establish a safe genetic diagnosis. It is therefore beholden upon those who generate and interpret genetic test results to allow access to relevant data as widely and as openly as possible, by depositing the data into appropriate databases and making it available to others to access. Numerous databases exist with aggregated genetic information, and although they differ in their deposition requirements and governance structures, ensuring interoperability between them will prevent information silos and ensure longer-term sustainability.

Despite the overwhelming benefits of genetic variant sharing, and paucity of proven harms, there remain ambiguities over the level of consent required for deposition of individual genetic variants to open access databases. We propose that consent should not be required for widespread, open sharing of individual pseudonymised genetic variants linked with high-level phenotypes, and that sharing such data should become standard practice in genomic medicine. We also recommend that richer case-level phenotypic detail is shared within healthcare systems to facilitate robust diagnosis and that consent is routinely sought at the time of diagnosis to share such data openly. Ultimately, both the promise and the safety of genomic medicine depends on our ability to share.

Data availability

No data are associated with this article

Grant information

This work was supported by the Wellcome Transforming Genomic Medicine Initiative [200990].

The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Acknowledgments

The authors wish to thank Fiona Cunningham, Ewan Birney, Matthew Hurles, David FitzPatrick, Graeme Black and Patrick Chinnery for helpful comments and input on this manuscript.

References

2. Turnbull C, Scott RH, Thomas E, et al.: The 100 000 Genomes Project: bringing whole genome sequencing to the NHS. BMJ. 2016; 353: k1687. Publisher Abstract | Publisher Full Text


Open Peer Review

Current Peer Review Status: ? ✓

Version 1

Reviewer Report 11 March 2019

https://doi.org/10.21956/wellcomeopenres.16463.r34792

© 2019 Martin C et al. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Christa L Martin
Autism & Developmental Medicine Institute (ADMI), Geisinger Health System, Danville, PA, USA

Erin Rooney Riggs
Autism & Developmental Medicine Institute, Geisinger Health System, Lewisburg, PA, USA

Heidi Rehm
Massachusetts General Hospital, Boston, MA, USA

Answers to the questions:

1. Is the rationale for the Open Letter provided in sufficient detail?

The authors state that the rationale for their recommendation to share genomic variants with limited clinical information is to encourage consistency and transparency among the genetics community, and to bolster the practice of genomic medicine, making it more beneficial to patients.

2. Does the article adequately reference differing views and opinions?

The article does provide a section outlining “perceived harms of sharing genetic variant data,” effectively outlining commonly proposed concerns about genomic data sharing. To be clear, our group, the Clinical Genome Resource (ClinGen), has also published articles with similar recommendations to those put forth by these authors on genomic data sharing, and we agree with the concepts presented in this manuscript. However, we are aware of at least one other “differing” opinion that was not represented here: the opinion that public data sharing highlights discordance in variant interpretation and is potentially confusing for clinical users 1,2. Our group believes that exposing discordant classifications between laboratories is actually a benefit to data sharing, allowing laboratories to see where they differ and work together towards concordance 3,4,5.

3. Are all factual statements correct, and are statements and arguments made adequately supported by citations?

In our general response, we note one place where we thought a factual statement was not completely
accurate. The authors state: “...US-based ClinVar is perhaps the leading genetic variant deposition database... but most variants have only very limited or no clinical information and no supporting evidence associated with them.” While it is true that most entries do not contain patient data, the majority (62%) of the more than 170,000 pathogenic/likely pathogenic variants in ClinVar have supporting evidence, either as written evidence summaries and/or PubMed citations.

There were also some other places in the document where additional context would be useful (for example, in order for the reader to understand the nuances between variant-level and case-level data sharing or to further explain the difference between the disease upon which a claim of variant pathogenicity was made and the phenotypic features presenting in an individual patient). These issues are noted in our general response.

4. Is the Open Letter written in accessible language?

Yes

5. Where applicable, are recommendations and next steps explained clearly for others to follow?

Providing more clarity would be helpful regarding “next steps” for readers to follow, particularly in regards to where variant information could be submitted.

General report:
Wright and colleagues have written an open letter to address genomic variant sharing. It is a thorough and excellent accounting of the rationale for this type of data sharing and we commend the authors for taking the time to thoughtfully review and provide guidance on this important topic. We have a few suggestions and some minor edits that could strengthen the article and provide additional guidance to the community.

Higher level comments and suggestions:
In the first section under “Recommendations” the first recommendation suggests only sharing “plausibly causal” genetic variants. We think this recommendation is insufficient and strongly encourage this guidance to include sharing of all variants that have been reviewed. The literature and databases are currently littered with false claims of causality/pathogenicity. It is critically important that we also share evidence on variants that are deemed benign or uncertain, or, at the case-level, deemed non-causal. Over three-quarters of ClinVar’s content is made up of variants classified as benign, likely benign, or variant of uncertain significance (VUS) and this data has been enormously useful to counter many of the false claims of pathogenicity from the literature.

In addition, there is a bit of conflating of the concept of variant-level versus case-level interpretation and it would be useful to better separate these concepts in the paper. We have previously defined “variant-level” information as the aggregation of all evidence and observations to define the pathogenicity of a variant (i.e., its capacity to cause disease). This may include evidence from a current case under investigation, but also takes into account all prior available data. However, whether a given variant is actually causal for the symptoms in a given patient is best called case-level interpretation and involves additional factors such as penetrance, a phenotype match with the relevant gene, and allelic information (e.g., recessive disease requires two alleles).

Related to this issue, we recommend in the section that outlines five specific data elements and references our prior publication, that items 2 and 4 be swapped to start with the variant level claim and then include the patient phenotype as part of the supporting evidence. This approach is more in line with
variant-level data sharing and our referenced publication, which should be distinguished from case-level sharing, which is also important, but requires additional considerations as the authors have pointed out. Similarly, the variant claim (e.g., pathogenic, benign) should be asserted against a disease, not the patient’s clinical features, which should be left to the case-level interpretation step. We have made suggested edits below to data elements 3 and 4 to better clarify these points (additions in bold, deletions indicate by strikethrough):

“Specific data elements for sharing individual genetic variants have been outlined previously and include (see Table 1):
1. a standardised genetic description of the variant(s), including Human Genome Variation Society (HGVS) nomenclature and genomic coordinates of the variant;
2. the clinical significance and summary of evidence upon which that assertion was based;
3. the disease and inheritance pattern of the disease (e.g. dominant/recessive) upon which the clinical significance is asserted;
4. a standardised clinical description of any of the clinical features in the patient(s) that are included as supporting observations for the variant assertion, using appropriately controlled vocabulary/ontology; and
5. a cryptic link to the laboratory or clinical service that submitted the data (to enable further information to be requested and avoid data duplication).”

Next, the authors state, “We recommend a cryptic link to the individual case-level data is maintained in a de-identified fashion via the laboratory or clinical service that submitted the data, that may obscure its geographical location by deposition via another platform, to enable clinical follow-up if needed.” We think this topic requires further consideration of the benefits and drawbacks of obscuring the submitter’s location. For most laboratories that perform a large volume of testing and receive samples from geographically diverse locations, it seems unnecessary to obscure the geographical location of the laboratory and data; indeed, the geographical location of the laboratory is easily discernible given that individual variants are attributed to specific laboratories in databases, such as ClinVar. ClinVar has operated with transparency to the submitter and their location without harm for several years now. To the contrary, it can be helpful to recognize the potential for data duplication, which is not uncommon. We would suggest a more nuanced discussion of this topic. For example, one may consider obscuring geographical location only in instances where the population is small or geographically isolated.

Finally, it would be useful if the authors gave more concrete suggestions for where laboratories should submit their classified variants today. Do the authors support that direct submission to ClinVar is one recommended option? The authors describe both ClinVar and DECIPHER, but it is unclear what their recommendation would be. Given the momentum that ClinVar has achieved, it seems important that wherever the variant classifications are initially generated and stored, that they also be easily submitted to ClinVar. DECIPHER does have the advantage of a richer connection to case-level data. If DECIPHER took on a role as an additional site of primary variant deposition (not clear if it accepts individual submitted variant interpretations), we would assume the authors would agree that it would still be important for DECIPHER to facilitate submission to ClinVar on behalf of its users, in the same way that DECIPHER is able to fully consume ClinVar data. Would this be a second recommended option? More detail around any recommendations and/or future plans would likely be useful to readers.

Minor suggestions and edits:

In the abstract the authors state: “We therefore recommend that widespread sharing of a small number of individual genetic variants associated with limited clinical information should become standard practice in
genomic medicine.” We assume the authors mean a small number “per individual” but we read it as stating that each “source/laboratory” should only share a small amount of data, in general. We suggest deleting “a small number of” and saving the nuance of per individual issue for later in the paper. Alternatively, you could reword the statement to read: “We therefore recommend that widespread sharing of a small number of an individual’s genetic variants associated with limited clinical information should become standard practice in genomic medicine.” This same issue occurs in the second paragraph of the section “Finding a balance”. In the sentence “…sharing a tiny subset of..” we suggest adding “an individual’s” after “of”.

In the last sentence of the abstract the authors state “For additional case-level detail about individual patients or more extensive genomic information, which is often essential for clinical interpretation, it may be more appropriate to use a controlled-access model for data sharing…..”. We fear this could be implied as abandoning the core suggestion if one wants to also share case-level data and therefore we suggest clarifying by adding “this additional” so the sentence reads “…it may be more appropriate to use a controlled-access model for this additional data sharing…..”

For the second Recommendation “A single genetic variant is not personally identifiable information; however, it is good practice to maintain a cryptic link to the laboratory or clinical service that shared the genetic data so that clinical follow-up remains possible should knowledge of the implications of a variant change.” We suggest adding “or to combine data to build evidence”. In our experience, many variants change classifications once labs bring their evidence/observations together, but a source for contact is needed to communicate and bring the data together.

Another good example of patient benefit, and avoidance of harm, from data sharing is Grant et al, referenced below, in case you would like to cite.

The authors state: “....US-based ClinVar is perhaps the leading genetic variant deposition database……but most variants have only very limited or no clinical information and no supporting evidence associated with them.” This statement is not completely accurate. While it is true that most entries do not contain patient data, the majority (62%) of the >170,000 pathogenic/likely pathogenic variants have supporting evidence, either as a written evidence summary and/or PubMed citations.

References


**Is the rationale for the Open Letter provided in sufficient detail?**
Yes

**Does the article adequately reference differing views and opinions?**
Partly

**Are all factual statements correct, and are statements and arguments made adequately supported by citations?**
Partly

**Is the Open Letter written in accessible language?**
Yes

**Where applicable, are recommendations and next steps explained clearly for others to follow?**
Partly

**Competing Interests:** The reviewers are investigators who receive funding from NIH/NHGRI for the Clinical Genome (ClinGen) Resource project (U41HG006834), an initiative with a focus on data sharing. In addition, several authors on the manuscript under review participate in ClinGen working groups, and ClinGen and DECIPHER co-host an annual scientific conference.

**Reviewer Expertise:** Genomic variant curation and clinical interpretation, broad data sharing of variant and phenotypic data

We confirm that we have read this submission and believe that we have an appropriate level of expertise to confirm that it is of an acceptable scientific standard.

Reviewer Report 08 March 2019

https://doi.org/10.21956/wellcomeopenres.16463.r34794

© 2019 Matthijs G. This is an open access peer review report distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.
Correct genomic variant interpretation and classification are an important and often complex issue, not only in research, but also, and even more critically, in genetic diagnostics and clinical care. Variant interpretation relies on different criteria and different in silico tools are available. The authors rightly argue that sharing of genomic variants is essential for improving and facilitating variant interpretation. Indeed, patients will strongly benefit from open and well-managed databases. Databases may be federated, as long as a swift and controlled exchange of information is available.

Clearly, variant sharing is more than a clinical or technical issue, it is a societal issue: citizen – and not just patients – should be informed about the necessity and value of variant sharing. They should be convinced that variant sharing is essential and safe. It is a matter of solidarity and mutual interest to share data as broadly as possible. The principle of proportionality, in relation to potential harm, is certainly rightly applicable to variant sharing.

Comment on the abstract and beyond:
- The authors use ‘de-identified’ and ‘pseudonomised’, ‘cryptic link’ and a few other descriptions. It would be good to select the best term or definition and explain it to the readership.
- It is unclear what is meant in the abstract with “a small number of …”. Small is hard to define.
- “Information robustly linking genetic variants with specific conditions is fundamental biological knowledge.” This is a significant statement that should be explored and explained in more detail, especially if the statement adds that it “should not require consent…”.

On the recommendations:
- In recommendation 1. The ‘small number’ from the abstract is not well reflected (or vice versa). What are ‘high level’ phenotypes, and why would sharing be limited to these?
- Recommendation states no consent in 1. and explicit consent in 3. This dichotomy is not presented in the Abstract. Again, the definition of ‘small’ is crucial, in all instances of policy, defining a ‘cut-off’ is a tricky thing.
- In general, what about sharing genomic variant that are excluded from disease, i.e. definitely not linked to the disease? The best example is in trans with a known dominant, pathogenic mutation. That information is equally useful.

On the Advantages:

For 2. Clearly, individuals may be identified in data bases by the genotype, for inclusion in clinical trials. With whom shall the data be shared? Companies? What would be the conditions? Who shall be the custodian? How to warrant and permit access? It would be nice to elaborate a bit on this. It is another aspect of variant sharing, that is not covered under the umbrella of variant interpretation.

For 3. The moral duty to help has been turned into a legal obligation in France. It may be interesting to cite this, as it is an example or situation that may pop up in other countries.

For documentation of the situation in France, please visit the following sites:
https://www.legifrance.gouv.fr/affichCodeArticle.do?cidTexte=LEGITEXT000006072665&idArticle=LEGIA
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT0000027592003&categorieLien=id
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000029921462
https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT0000027592025&categorieLien=id

For 4. How shall data be linked to natural history of disease, or vice versa?
On the Disadvantages:

- Ref 29 is not tightly linked to the issue of the proposed international sharing of data. Are there other cases/references?
- There are other, early papers on re-classification, e.g. Piton et al. 2013.
- It is probably worthwhile to mention that some international databases are ‘contaminated’ i.e. contain a wrong and erroneous variant classification. So the curation is essential. Variant database should explicitly how the data is collected and managed. In the diagnostic arena, there is a consensus that HGMD data should be explicitly double-checked.
- The authors also mention the issue of private databases. It is unfortunate indeed that genetic and genomic analyses that are performed in often commercial laboratories do not make it to the public databases. Several large laboratories, mostly in the US, are committed to sharing data, like for instance via ClinVar. However, bad examples do exist as well, and have been denounced early on. References could be added, e.g. Conley et al 2014.
- Also of note is that several databases, that were originally open, have been acquired by commercial companies. The latter has been favourable for their survival, however, the licencing fees are often prohibiting individual laboratories to obtain access. Equally, some companies offer access to their own clinical diagnostic databases, but again, the prices are mostly prohibitive. The data in private databases, especially these that are well curated, may be considered as having a value – as a result of intellectual or other efforts to generate good data – and thus come with a price. How to deal with this?
- In parallel, the public laboratories have not been very active in submitting variants. What kind of incentive would be needed to promote data sharing?
- The statement on informed consent hints at an important shift in the policy of Decipher to request consent. This policy was reportedly very strict in the early days. The position statement pleads for a relaxed (or no) requirement for a written consent. It would be interesting to read how and why the policy has evolved so significantly.

Finding a balance:

- What is the link between the text and Table 1. Table 1 does not list all the elements that are listed in the text. It would be good to explain to the reader what the aim of Table 1 is.
- Open versus controlled access: open access shall best be promoted, given the large number of labs that will either submit or consult.
- Open access databases are not necessarily free. Are there any other incentives to urge (diagnostic or research) laboratories to share variants? The latter are invited to submit in relation to publication, the former? Linking it to reimbursement of the test would be an option for laboratories operating in a ‘fee for service’ (public) health system, but would not be useful for private billing.
- What about using a model of clearing houses, to offer an incentive for submission? At some moment, funding and/or a financial model for maintaining the databases will be necessary.
- Page 6: LOVD shall be mentioned, as it fulfils the criteria listed in the text.

On information silos:

- It would be good to give a brief description and view point on how silos could be broken down or avoided. What about a model of federated databases?

References

Publisher Full Text

Is the rationale for the Open Letter provided in sufficient detail?
Yes

Does the article adequately reference differing views and opinions?
Partly

Are all factual statements correct, and are statements and arguments made adequately supported by citations?
Yes

Is the Open Letter written in accessible language?
Yes

Where applicable, are recommendations and next steps explained clearly for others to follow?
Yes

**Competing Interests:** No competing interests were disclosed.

**Reviewer Expertise:** Molecular genetics, rare diseases, NGS diagnostics, societal issues of genomic medicine

I confirm that I have read this submission and believe that I have an appropriate level of expertise to confirm that it is of an acceptable scientific standard, however I have significant reservations, as outlined above.